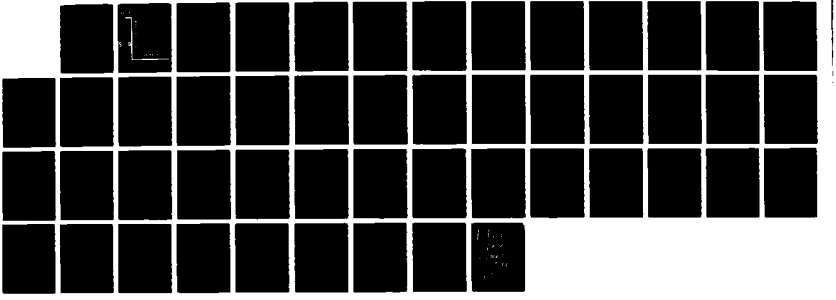

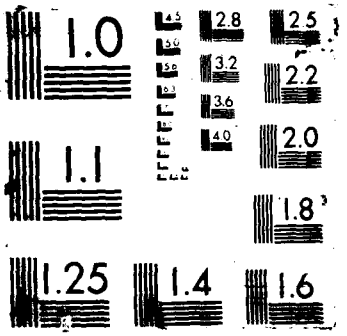
1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0

AFHRL-TP-87-58

(2)

AIR FORCE HUMAN RESOURCES LABORATORY

JOB PERFORMANCE MEASUREMENT: TOPICS IN THE PERFORMANCE MEASUREMENT OF AIR FORCE ENLISTED PERSONNEL

M. Suzanne Lipscomb
Jerry W. Hedge

TRAINING SYSTEMS DIVISION
Brooks Air Force Base, Texas 78235-5601

June 1988

Interim Technical Paper for Period March - July 1987

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

HENDRICK W. RUCK, Technical Advisor
Training Systems Division

GENE A. BERRY, Colonel, USAF
Chief, Training Systems Division

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No 0704-0183

1a REPORT SECURITY CLASSIFICATION Unclassified	1b RESTRICTIVE MARKINGS
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b DECLASSIFICATION/DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBER(S) AFHRL-TP-87-58	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a NAME OF PERFORMING ORGANIZATION Training Systems Division 6b OFFICE SYMBOL (If applicable) AFHRL/IDE	7a NAME OF MONITORING ORGANIZATION
6c ADDRESS (City, State, and ZIP Code) Air Force Human Resources Laboratory Brooks Air Force Base, Texas 78235-5601	7b ADDRESS (City, State, and ZIP Code)
8a NAME OF FUNDING/SPONSORING ORGANIZATION Air Force Human Resources Laboratory 8b OFFICE SYMBOL (If applicable) HQ AFHRL	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c ADDRESS (City, State, and ZIP Code) Brooks Air Force Base, Texas 78235-5601	10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO: 62703F; PROJECT NO.: 7734; TASK NO: 08; WORK UNIT ACCESSION NO: 22

11 TITLE (Include Security Classification)
Job Performance Measurement: Topics in the Performance Measurement of Air Force Enlisted Personnel

12 PERSONAL AUTHOR(S)
Lipscomb, M.S.; Hedge, J.W.

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
Interim	FROM Mar 87 TO Jul 87	June 1988	52

16 SUPPLEMENTARY NOTATION
→ (Armed Services Vocational Aptitude Battery)

17 COSATI CODES		
FIELD	GROUP	SUB-GROUP
05	08	
05	09	

18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
ASVAB, job performance, performance measurement, Air Force training, technology transfer, test content selection, training evaluation, SDW, aptitude tests.

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

In the early 1980's, the Air Force initiated a long-term effort to develop a performance measurement technology for accurately measuring the performance of enlisted personnel. Requests from operational military and civilian program managers in the manpower, personnel, and training (MPT) communities for criteria to evaluate training and selection programs, and requests from the MPT research community for criteria to validate their R&D projects, energized research in the performance measurement domain. A Congressional mandate to test the feasibility of linking enlistment standards to job performance added impetus to this effort. This document contains a series of five papers which discuss current work in this area being done by the Air Force Human Resources Laboratory. Detailed descriptions of AFHRL's strategy for test content selection, work sample testing, data analysis, training evaluation, and technology transfer are presented and discussed.

Keywords

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	Unclassified
22a NAME OF RESPONSIBLE INDIVIDUAL Nancy J. Allin, Chief, STINFO Office	22b TELEPHONE (Include Area Code) (512) 536-3877 22c OFFICE SYMBOL AFHRL/TSR

DD Form 1473, JUN 86 *Previous editions are obsolete*

SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

AFHRL Technical Paper 87-58

June 1988

JOB PERFORMANCE MEASUREMENT: TOPICS IN THE PERFORMANCE MEASUREMENT OF AIR FORCE ENLISTED PERSONNEL

M. Suzanne Lipscomb
Jerry W. Hedge

TRAINING SYSTEMS DIVISION
Brooks Air Force Base, Texas 78235-5601

Reviewed and submitted for publication by

Jerry W. Hedge
Chief, Skills Development Branch
Training Systems Division

This publication is primarily a working paper. It is published solely to document work performed.

SUMMARY

This document contains a series of papers presented at a Department of Defense/Educational Testing Service conference held in San Diego, California in March 1987. As such, it describes ongoing research and development (R&D) within the Air Force Human Resources Laboratory's Job Performance Measurement Project. Papers on test content selection, work sample testing, predictive efficiency of the Armed Services Vocational Aptitude Battery, training evaluation, and transfer of these technologies to users within the DoD user community demonstrate both the breadth of R&D and the broad applicability of this performance measurement technology.

PREFACE

Active research and development in the area of job performance began in the early 1980's across the four Services in response to a Congressional mandate and requests from the Services' user communities. In March 1987, the Department of Defense and the Educational Testing Service hosted a conference on Job Performance Measurement Technologies to highlight Service efforts and to serve as a forum for discussion among the Services themselves and public and private agencies engaged in similar activities. The papers presented here describe the Air Force Job Performance Measurement effort.

TABLE OF CONTENTS

Table of Contents (Continued)

Table of Contents (Concluded)

LIST OF FIGURES

LIST OF TABLES

JOB PERFORMANCE MEASUREMENT:
TOPICS IN THE PERFORMANCE MEASUREMENT
OF AIR FORCE ENLISTED PERSONNEL

I. TEST CONTENT SELECTION

M. Suzanne Lipscomb
Air Force Human Resources Laboratory

Terry L. Dickinson
Old Dominion University

In the development of any test, it is rarely possible to construct and administer items which completely exhaust the domain of content to be measured. Time and other practical considerations constrain what can be covered in any given testing situation. Thus, unless the content to be measured is very narrowly defined, it is necessary to rely on samples to generalize to the domain and, ultimately, the universe. This is particularly true in the case of hands-on work sample testing, where the number of tasks that can be covered is restricted.

The quality of the generalizations or inferences made from test scores is directly related to the quality of the definition and the sampling of the domain. In order to make valid generalizations, the domain must be well defined and the sampling must be relevant and representative. This requirement for the test to represent the larger domain also extends to the selection of types of items, item quality, and the administration and scoring procedures used. The critical question becomes, "Does a person's score on this test reflect his/her standing on the entire domain of interest?" This, in turn, leads to the question of test validity.

Defined, the validity of a test concerns how well a test measures what it purports to measure (Allen & Yen, 1979; Anastasi, 1982). Thus, validity refers to the accuracy of predictions or inferences made from test scores, taking into consideration the particular use of the test (Cronbach, 1971).

Though there have been calls for a more unified view of the validation process (Landy, 1986), procedures for determining the validity of a test are often classified within three general categories: content, criterion-related, and construct validity. These three types of validation procedures are interrelated, with each addressing a specific aspect of the test and the interpretation of scores on the test. Broadly defined, content validity refers to the extent to which the content of the test represents the behavioral domain to be measured. Criterion-related validity reflects the effectiveness of a test in predicting a person's behavior in a specified situation, either concurrently with the test or in the future. Construct validity is concerned with the extent to which a test measures a theoretical construct or trait. It is evaluated by investigating the degree to which certain explanatory concepts account for performance on the test (Cronbach, 1971).

Quality tests are constructed with the goal of validity in mind. It is the aim of the test constructor to develop a test which measures what he/she has set out to measure, whether it is a trait, aptitude, or achievement level. The first step in the test development process is the specification of what is to be measured. This can be viewed as a process of identifying the performance universe which encompasses all possible behaviors relevant to the measurement goal. The content domain identifies a portion of the content universe for the purposes of testing. The test content universe, which can then be specified at least theoretically, consists of all possible test items that can be developed for the content domain, as well as conditions of testing and scoring procedures. From this test content universe, a sample of items is taken

which defines the actual specifications for test construction. This constitutes the test content domain from which the test is constructed (Guion, 1979).

It is this process of specifying the sample of what is to be measured and how it is to be measured that ultimately determines the validity of the measure. Although this process is usually the focus of content validity evaluation, it also has direct impact on the criterion-related and construct validity of the test. Misspecification of any of the areas of interest from the content universe to the test content universe, or an inappropriate sampling procedure, may result in the measurement of something other than what was intended.

It is this process of domain specification and sampling that is the focus of this paper. The strategy used to develop Air Force Walk-Through Performance Tests will be described, and issues relating to the use of such a strategy will be discussed.

THE AIR FORCE DOMAIN SPECIFICATION AND SAMPLING PLAN

When developing task-based job performance measures, it is impractical to assess performance on the universe of tasks within most Air Force specialties (AFSs). No individual performs all of the tasks in any specialty, and no individual performs an average job in most specialties. Rather, the tasks of a specialty are distributed by management action to individuals in consistent ways so as to cluster into a variety of types of jobs. These clusters are based on the co-performance of tasks and the variations in mission, equipment, or management in any given locale. This variance of jobs within AFSs is an exceedingly important phenomenon, as it impacts on how the specialty is organized in the personnel system, the aptitudes required, the training provided, and the way individuals can be utilized in the workplace (Mitchell & Driskill, 1979). This variance in Air Force jobs is of concern to Air Force managers and is one of the major issues of study in the occupational analysis program (AF Regulation 35-2, Occupational Analysis Program). Data on most AFSs indicate that the classification structure of the Air Force is highly dynamic, with frequent reallocation of tasks among specialties. In addition, while there may be some common tasks performed by a majority of individuals in a specialty, most of the tasks are performed only by members of the various job types within the specialty.

It is necessary, therefore, to rely on samples of performance that are both useful for differentiating between good and poor performers and representative of the performance domain. Differentiating good and poor performers can be accomplished by assessing job incumbent performance on tasks with a range of difficulty. In addition to identifying tasks with a range of difficulty, selecting tasks that adequately represent the total specialty domain is necessary to make inferences about performance from a sample of specific tasks. If the specialty domain is adequa ly represented by the tasks selected, the task-based measurement system can be considered content valid. Unlike other aspects of test validity, the content validity of a measurement procedure is not a correlational process but an evaluation of adequacy and representativeness using rational judgments. Lennon (1956) stated that three assumptions underlie the use of content validity: (a) The area of concern to the user can be conceived as a meaningful, definable universe; (b) a sample can be drawn from the universe in some purposeful, meaningful fashion; and (c) the sample and the sampling process can be defined with sufficient precision to enable the user to judge how adequately the sample typifies the universe.

Given the information available in the Air Force Occupational Research Data Base, these three assumptions can be met; thus, the issue of content validity can be addressed. The universe can be defined as the set of tasks for an AFS as detailed by the occupational survey task list. The sample can be drawn in a meaningful fashion based on the task-level occupational survey data available. Finally, the sampling process can be defined with precision using a task sampling plan which will allow a judgment to be made as to the adequacy of the sample.

The task sampling plan consists of a procedural set of guidelines which: (a) specify the task clusters (i.e., jobs) of interest, (b) establish the level of measurement specificity, and (c) determine the proportional weighting (importance) of the work activities identified. These guidelines assure objectivity, replicability, and comparability of efforts to develop measures which detect meaningful differences in performance, and their use is illustrated below for the first AFS investigated by the Air Force for the Joint-Service Job Performance Measurement project, Jet Engine Mechanic (AFS 426X2).

Task Selection Procedural Guidelines

Defining the Job Domain

For most AFSs, the Air Force has a wealth of information sources which give a comprehensive picture of the work domain; for example, AFS entrance requirements and a general specialty description (AFR 39-1, Airman Classification Regulation); AFS training requirements (AFM 50-5, USAF Formal Schools Catalog and Specialty Training Standards); and occupational survey data.

Occupational survey data include: the percentage of incumbents who report performing each task, the amount of time incumbents report performing the task relative to other tasks, subject-matter experts' (SMEs) judgments of the relative time required to learn to perform tasks (i.e., task difficulty), and the relative importance of training for each task (i.e., training emphasis). Occupational survey data and training information, which cover the full scope of tasks performed by incumbents in an AFS, were applied to the development of the task-based performance measures. Because occupational survey data are the most detailed and comprehensive, they were used to define the work domain. The other sources provided complementary information.

The goals of the Air Force Job Performance Measurement Project are to assess specific job competencies required within a specialty, and general competencies applicable across AFSs. These two types of measures require four levels of measurement specificity: Air Force-wide, specialty-wide, duty-core, and incumbent-unique measures. Since the focus of this paper is on selecting tasks required to measure individuals' competence within an AFS, only the latter three levels of measurement specificity are described.

To include an adequate representation for each of these three levels of measurement specificity, tasks within an AFS must be categorized accordingly. That is, tasks must be categorized into those performed throughout the specialty (i.e., specialty-wide), those specific to certain duties within an AFS (i.e., duty-core), and those uniquely performed by incumbents in certain job types (i.e., incumbent-unique).

The occupational survey task inventory was used to define the work domain and categorize tasks. Because task performance is often specific to equipment or work centers, tasks associated with equipment or work centers were used to identify the duty-core domain. Finally, tasks associated with specific job types defined by the occupational analysis were used to delineate the incumbent-unique domain. Since it would be impractical to cover adequately all duty areas and job types within heterogeneous AFSs, those most representative of the work performed were selected. That is, duty areas and job types which have the largest percentage of personnel were chosen.

Selecting Tasks Representative of the Job Domain

The procedures for sampling tasks representative of the three levels of measurement specificity are outlined in the following paragraphs, along with the rationale for these procedures.

For each task domain, the number of tasks selected was based on a judgment of the number of performance measures required to give an adequate sample and conformed to a total testing time of no more than 8 hours for all measures. This time limit was the maximum time feasible to keep an airman from his/her unit. Within this timeframe, individuals were assessed on specialty-wide, duty-core, and incumbent-unique tasks.

Phase I. Selection of Specialty-Wide Tasks

Step 1. Select all tasks which are included in the Plan of Instruction (POI) for initial AFS training or, if not in the POI, are performed by at least 30% of the first-term incumbents, 1-48 months total active Federal military service (TAFMS).

This reduced the task pool to those tasks deemed important enough for training or those performed by a substantial number of first-term airmen across the AFS. (The 30% cutoff value may be varied by specialty according to the number of tasks performed by first-termers in that specialty.)

Step 2. Cluster tasks selected in Step 1 based on one or more of the following: (a) factor or cluster analysis of co-performance data, (b) Specialty Knowledge Test outline, (c) Specialty Training Standard outline, or (d) occupational survey duty outline. Each of these is a means of organizing the pool of tasks into performance/knowledge areas based on occupational information. All are similar in results; thus, the selection of the clustering strategy should be based on a judgment as to which is cost effective and well suited to the development of performance measures for a specific AFS.

Step 3. Weight each task cluster to reflect its relative importance to the overall performance of first-term airmen within the specialty. Possible sources for weighting clusters include the following: (a) Specialty Knowledge Test outline weights, (b) Specialty Training Standard proficiency level requirements, (c) SME judgments of relative importance, and (d) weights derived from training emphasis ratings (i.e., SME judgments of the extent to which training is required for tasks) and percent time spent ratings (i.e., incumbent judgments of the relative time spent performing tasks). These latter weights could be derived as the product of the mean training emphasis rating and the cumulative time spent performing tasks in a cluster.

Step 4. Determine the number of tasks to be selected from each cluster to reflect the assigned weights. Total the cluster weights, and divide each cluster weight by the total to get its relative percentage of importance. Multiply each cluster percentage by the total possible tasks to determine the number of tasks to be selected.

Step 5. Within each cluster, randomly select the number of tasks determined in Step 4 to reflect a range of learning/task difficulty by: (a) ranking the tasks on task difficulty, (b) dividing the ranked list into quartiles, (c) selecting 40% of the tasks from the fourth quartile, (d) selecting 30% from the third quartile, (e) selecting 20% from the second quartile, (f) selecting 10% from the first quartile, and (g) repeating for each cluster. (It is important to sample tasks covering a range of difficulty so incumbent performance assessment will reflect the rank-ordering of people of varying levels of job competence. The sampling is more heavily weighted on the more difficult tasks to reflect the aptitude requirements of the specialty and where more performance variation should occur.)

Step 6. Review the tasks identified in Step 5 to determine if they can be measured by either the hands-on or interview components of Walk-Through Performance Testing (WTPT). Return any task found to be unsuitable for WTPT to the task pool. If possible, randomly select a replacement task from the same task difficulty quartile. Document why the original task was unsuitable.

(The ability to assess performance through observation/interview procedures (WTPT) is a prerequisite to final task selection because performance measures obtained via these high-fidelity techniques will be the benchmarks against which surrogate measures are compared.)

Phase II. Selection of Duty-Core Tasks

Because the performance domain for a duty area (e.g., a specific engine type or work center) is much less broad than for the entire specialty, fewer tasks are needed for an adequate sample. Also, because tasks selected for one duty area may be performed in another area, tasks can be selected for more than one duty area. However, since tasks selected in Phase I for specialty-wide measures will be used to assess all incumbents, they should not be used to develop duty-core measures. The following steps apply for each duty area.

Step 1. Select all tasks performed by at least 40% of the first-term airmen identified as performing the duty in question (as noted earlier, this cutoff may vary according to number of tasks performed by first-termers) and not utilized in Phase I. (Within each duty area, a higher proportion of incumbents performing tasks can be used as a basis for identifying tasks to be assessed because the performance domain is more narrowly defined than across the entire specialty.)

Step 2. From the tasks identified in Step 1, select the total number of tasks to reflect a range of learning/task difficulty by repeating Phase I, Step 5.

Step 3. Repeat Phase I, Step 6.

Phase III. Selection of Incumbent-Unique Tasks

Because the performance domain for each job type is less broad than for the entire specialty, fewer tasks are needed to provide an adequate sample. Also, because the tasks selected for a job type may be applicable to another job type, tasks selected for one may be used for another. However, tasks selected in Phases I or II should not be used to develop incumbent-unique measures. The following steps apply for each job type.

Step 1. Select all tasks performed by 50% or more of the incumbents in the incumbent-unique group and not utilized in Phases I or II. (Again, as the job domain becomes more specific, it is possible to select tasks performed by a higher proportion of incumbents. In addition, the cutoff may vary by number of tasks performed by first-termers.)

Step 2. From the tasks identified in Step 1, select the total number of tasks to reflect a range of learning/task difficulty by repeating Phase I, Step 5.

Step 3. Repeat Phase I, Step 6.

Review and Approval of Task Sample

A description of the application of these task sampling procedures and the tasks selected for each AFS was reviewed by appropriate AFS functional managers and technical training representatives, who provided feedback concerning the adequacy of the tasks selected. Reviewers examined the task sample to ensure that work performed by first-term airmen and critical wartime requirements were well represented. Approval of the task sample by these policy makers should increase the acceptance and utilization of the resulting job performance measures.

APPLICATION TO THE JET ENGINE MECHANIC SPECIALTY (AFS 426X2)

Before the sampling plan could be applied to the Jet Engine Mechanic Specialty, duty areas and incumbent-unique job types were identified. How these were chosen is described before illustrating employment of the task selection plan.

Defining the Job Domain

Duty Areas Selected. Duty areas were selected based on engine type maintained. An inspection of the occupational survey data revealed that 20%, 18%, and 17% of AFS 426X2 first-term airmen performed maintenance tasks on J-57, J-79, and TF-33 engines, respectively. Since these percentages were the highest among the nine engine types maintained by AFS 426X2 personnel, these three engines were selected as being representative of equipment maintained by first-term jet engine mechanics.

Job Types Selected. The occupational survey data also revealed that the vast majority of first-term jet engine mechanics perform similar jobs (i.e., most airmen maintain similar engine accessory systems). The largest percentage of first-term incumbents in each major command spend the majority of their time performing general engine maintenance tasks in shop or on the flightline. As a result, these two functional areas were identified as being representative of AFS 426X2.

Phase I. Selecting Specialty-Core Tasks

Task Clustering. Tasks were clustered by occupational survey duty area because this grouping adequately reflected the work done in the specialty and was cost effective. Weights were computed based on the product of the mean training emphasis rating and the cumulative time spent performing tasks in a cluster. The following six task clusters received the weights indicated.

Cluster	Weight
Preparing and Maintaining Forms, Records and Reports	10
Performing Quality Control Functions	5
Performing Flightline Engine Maintenance Functions	10
Performing In-Shop Engine Maintenance Functions	20
Performing Test Cell Functions	5
Performing General Engine Maintenance Functions	50

Task Selection and Review. The remaining Phase I steps were followed, and 18 tasks were selected to reflect the weights outlined above. Ten tasks were selected for each engine type in Phase II and ten for each job type in Phase III. Selected tasks were reviewed by SMEs and unsuitable tasks deleted. New tasks were selected and reviewed and the tasks list finalized, giving a representative set of tasks upon which to develop performance measures. The main justifications for the task exclusions were:

Task not common to all engines (Phase I)
Task not common to all functional areas (Phases I and II)
Task unclear, too broad, complex, or trivial
Task not representative of functional area (Phase III)
Task performed differently on different aircraft (Phase I)
Overlapping or similar task

Task performed differently depending on how engine is shipped (air, rail, or truck) and its destination (depot, deployment)

Task performed differently depending on organizational unit (Examples: Some supervisors do not allow test cell personnel to transport engines. Strategic Air Command flightline personnel do not make entries on oil analysis request forms (DD Form 2026), but Military Airlift Command flightline personnel do make such entries.)

Equipment being changed within the year

In summary, a strategy for task selection was developed to sample tasks representative of the job content. This strategy was applied to the Jet Engine Mechanic Specialty (426X2), and the selected tasks were used to develop performance measures and standards.

CONCLUSION

The Air Force test content selection strategy combines expert judgment and stratified random sampling to select representative tasks. The input of SMEs, in the form of task factor data and review, helps focus the task selection process on areas of importance. This is significant given the limited number of tasks which can be covered in a hands-on-testing situation due to time and cost constraints. Thus, it is critical that only those tasks deemed most representative be considered for inclusion in the test. The random sampling aspect of the sampling procedure helps assure the generalizability of the sampled tasks. Thus, expert judgment and random sampling procedures are incorporated in the sampling strategy to develop a content-valid test consisting of an optimal set of tasks for use in performance testing.

There are several research questions that are raised in the consideration of the utility of a content selection plan. Questions that merit attention concern the choice and definition of parameters, the level of detail required, the degree of judgment involved, and the appropriateness and generalizability of one selection strategy across many different content areas. Research addressing these questions will have both practical and theoretical value. Such research is currently being planned by the Air Force.

II. WORK SAMPLE TESTING IN THE AIR FORCE JOB PERFORMANCE MEASUREMENT PROJECT

Jerry W. Hedge
M. Suzanne Lipscomb
Mark S. Teachout

Air Force Human Resources Laboratory

The Air Force Human Resources Laboratory (AFHRL) is conducting a large-scale effort to develop a measurement technology for systematically obtaining job performance data. The overall program of research calls for the development of criterion measures that allow for the collection of valid, accurate, and reliable job performance information. The chief aim of criterion development research is to identify the measurement technique that most faithfully represents relevant job behaviors. One approach that is considered to have high fidelity to the work environment is work sample testing. The focus of this paper is work sample testing in the Air Force Job Performance Measurement Project. The paper describes the work sample philosophy and developmental process used by the Air Force, and details test administrator qualifications, training, and data collection. Relevant data associated with hands-on and interview work samples will be presented, including a comparison of the two approaches.

AIR FORCE WORK SAMPLE TESTING

As noted by Wilson (1962), over the years the primary use of the work sample has been for personnel selection. However, this approach can be a valuable aid in the measurement of job proficiency. Typically, work sample tests involve an individual in performing a task or set of tasks relevant to that person's job and selected from the range of tasks performed by the job incumbent. The value of the work sample methodology lies in the fidelity with which the selected set of tasks allow measurement of an incumbent's job proficiency. This can also be a weakness of the technique. Unfortunately, work sample procedures normally identify critical tasks, discard those not practically measurable, and then simply allow the remainder to become the selected set of tasks to be measured. AFHRL's approach to work sample testing is an attempt to overcome this criterion deficiency problem.

Walk-Through Performance Testing

For the Air Force, hands-on testing is a particular problem because of the complexity and expense involved in performing many tasks. For example, some critical tasks cannot be measured by hands-on testing because these tasks tend to take too long to complete, require replacement of expensive parts, or risk possible damage to components. AFHRL has developed a new methodology to deal with these problems. This new approach, Walk-Through Performance Testing (WTPT), has as its foundation the work sample philosophy but attempts to expand the measurement of critical tasks to include those tasks not measured by hands-on testing, through the addition of an interview testing component (Hedge, 1984).

The hands-on component of the WTPT resembles a traditional hands-on work sample test designed to measure proficiency on a critical task. For example, one hands-on task may require an incumbent to install a starter on a jet engine. On the first page of the test administrator's manual, information is provided to the test administrator concerning: required testing time; tools, technical orders, and job guides; pertinent background information and required engine configuration; and test administrator instructions. While the starter is being installed, the

test administrator uses a checklist to indicate whether steps (e.g., lubricate the spline, index the position of the starter, and install the locking device) are performed correctly. Finally, a 5-point rating scale allows the test administrator to record an overall rating of proficiency on the task.

Many tasks are either too time-consuming, too costly, or too dangerous to measure by hands-on testing. Interview testing attempts to expand the content domain by measuring tasks that cannot be measured practically with the hands-on method. Interview testing requires the incumbent to explain the step-by-step procedures necessary for successful completion of the task. This allows the test administrator to assess an incumbent's proficiency-based strengths and weaknesses related to the performance of that task. For example, an interview item may test an incumbent's ability to determine the source of high oil consumption. Once again, on the first page of the adminstrator's manual, pertinent information is provided to the test administrator. While the incumbent is explaining how to perform the task, the test administrator uses a checklist to indicate whether the steps necessary for successful performance are correctly described. In addition, a 5-point overall proficiency rating is recorded by the administrator.

The interview testing is conducted at the work site in a "show-and-tell" fashion that allows the incumbent to "visually and verbally" describe how a step is to be accomplished (e.g., "that bolt is to be turned five revolutions" or "that component is to be lubricated prior to being assembled"). Thus, information on additional tasks can be collected along with hands-on information to provide a more thorough coverage of the content domain and a more accurate picture of an individual's job proficiency.

Task Sampling and Item Development

An extensive task sampling plan was developed for each job using information obtained from the Air Force's Occupational Survey Program (Lipscomb, 1984). This program maintains job content information for over 200 of the 250 specialties in the Air Force. Surveys for each of these 200 specialties are administered approximately every 4 years to keep the job content information current. Available information in each job content domain includes the tasks performed, the relative amount of time spent performing these tasks, and emphasis to be given to the tasks in training. This information was used to select tasks for developing WTPT components.

Visits were made to several Air Force bases in order to interview subject-matter experts (SMEs) about these tasks (Alba, Dickinson, & Lipscomb, 1985). Using the appropriate technical order, the SMEs were asked to describe the procedural steps involved in performing each task, whether procedures for performance on a task might differ by location, and whether the development of a hands-on test for a task was feasible (e.g., in terms of time, and safety of equipment and personnel). This information was used to delete those tasks not suitable for testing. Hands-on and interview tests were written for each of the remaining tasks. These tests were reviewed by SMEs, and based on their input, the tests were refined. Finally, the tests were field tested at several Air Force bases. To date, work sample tests have been constructed in four specialties: Jet Engine Mechanic (AFS 426X2), Information Systems Radio Operator (492X1), Air Traffic Control Operator (AFS 272X0), and Avionic Communications Specialist (AFS 328X0).

Test Administrator Training

Experts in each of the career fields under study were selected to serve as test administrators. These personnel consisted of active-duty senior-level Noncommissioned Officers (NCOs) provided by their Major Commands, or recently separated/retired former NCOs hired by the contractor. Because the test administrators were already experienced at performing their

particular jobs, training focused on the logistics of the effort, and improving their observation, recording, and interviewing skills. Logistics training focused on base arrangements, testing requirements, and potential problems. Test administrators were given an administrator's manual for use as a procedural guide in the field.

To emphasize observation and scoring skills, videotapes of job incumbents (actors) performing or describing how they would perform tasks allowed the test administrators to practice observing and recording performance for the hands-on and interview tests. Videotapes were constructed for multiple tasks for each AFS. Scenarios were generated by consulting SMEs as to where and how legitimate performance errors could be made within each task. This information was used to develop correct versions in which job incumbents performed task steps correctly, and incorrect versions containing realistic performance errors. One correct and several incorrect versions of task performance were videotaped.

Interview skills training utilized videotapes and role playing. This training emphasized correct and incorrect procedures to follow in gathering data through the interview, modeling of correct interviewer behaviors, and face-to-face interaction between the test administrator and job incumbent.

DATA COLLECTION AND ANALYSIS

Data collection was conducted in a standardized fashion for all four AFSs with work sample tests. Because data collection and analysis have been completed only on jet engine mechanics, collection and analysis details will be specific to that career field.

Data Collection

Three 3-man teams of test administrators tested first-term (13-48 months total active Federal military service) jet engine mechanics using the Air Force work sample methodology. Over an 8-hour period, 10 hands-on and 10 interview items were administered. Performance on 5 tasks was measured using both hands-on and interview items. Each of these 20 work sample items carried a point value of 10.

Instrument pretest was conducted by the nine test administrators at three Air Force bases in the continental United States (CONUS). Forty-two job incumbents were tested using the WTPT approach. Full-scale data collection occurred at 13 Air Force bases in the CONUS. The three teams of test administrators collected data from 255 incumbents who were randomly selected from the population of first-term mechanics at each base.

Data Analysis

Jet Engine Mechanic data analysis included pretest and full-scale data collection in order to address the following issues concerning work sample tests: (a) interrater reliability of test administrators; (b) test/retest reliability; (c) discriminability; (d) comparison of hands-on and interview methods; and (e) comparison of work samples with other relevant variables. For purposes of analysis, summary values for each instrument were derived by summing scores across work sample items. In this way, a hands-on summary score consisted of 100 points (ten 10-point items) and an interview summary score consisted of 100 points (ten 10-point items).

Reliability, Accuracy, and Descriptive Statistics

Interrater Reliability. Data from the three teams of test administrators were analyzed at three separate points in time. During the training workshop (mentioned earlier) videotaped task performances were shown to and scored by all test administrators. This allowed both an analysis of interrater consistency and an assessment of administrator accuracy (in comparison to the videotaped target scores). One month after the first training workshop, pretest data were collected from 14 job incumbents per administrator team. For each team, nine incumbents were assessed by a single administrator. The remaining incumbents were scored by the 3-member team, allowing an evaluation of interrater reliability. Two and one-half months after the first workshop, a retraining workshop was held, yielding data collection and analyses comparable to the first workshop.

Pairwise percent agreement indices were computed across the three teams of test administrators. The arithmetic averages of these indices for each team are reported in Table 1. The indices suggest that a high level of interrater reliability was obtained for the three teams at all points in time. In addition, interrater agreement tended to improve over time.

Table 1. Interrater Agreement for the Workshops and the Pretest

Team	Workshop 1	Pretest	Workshop 2
TF-33	74.36	78.71	83.55
J-57	84.93	89.96	89.95
J-79	76.41	85.81	84.10

Interrater Accuracy. Percent agreement accuracy indices were calculated for each administrator on common tasks performed in each workshop. These indices were computed between ratings and target scores. The averages of the indices for each team across tasks are reported in Table 2. Accuracy was quite high for all teams at both workshops.

Table 2. Percent Agreement Between Test Administrators and Target Scores for the Workshops

Team	Workshop 1	Workshop 2
TF-33	73.92	91.81
J-57	86.79	94.98
J-79	67.18	78.46

Test-Retest Reliability. Three pretest bases were revisited during full-scale data collection (approximately 2 months after pretest), and test administrators collected work sample data from all available job incumbents that participated in the pretest. Test-retest reliability estimates were computed for both hands-on and interview tests. For all hands-on tests, incumbents were rated consistently 78% of the time. Ratings with the interview were not as consistent, averaging a test-retest correlation of .56. This lower value can be attributed to the greater subjectivity of the interview procedure. Thus, these procedures take longer to stabilize than the more objective hands-on procedures.

Work Sample Descriptive Statistics. Ten hands-on work sample items comprised the hands-on work sample test. Thus, out of 100 points, the 255 jet engine mechanics scored an average of

73.01, with a standard deviation of 10.53. For the 10 interview work sample items that comprise the interview work sample test, incumbents scored 62.35 on the average, with a standard deviation of 12.53. These average scores suggest a test of moderate difficulty. A test with large numbers of incumbents scoring quite high or low would reduce the ability to discriminate between incumbents' performance. In fact, the variability in test scores is quite good, with the interview slightly superior to the hands-on test. Figures 1 and 2 show the distributions of scores for both the hands-on and interview work samples.

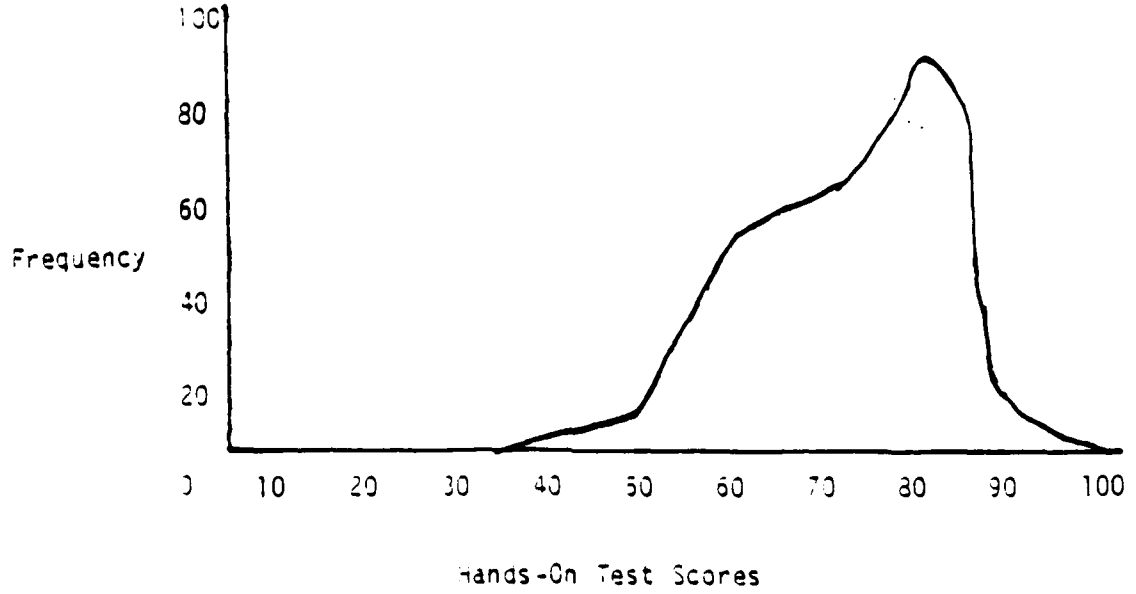

Figure 1. **Hands-On Work Sample Test Score Distribution.**

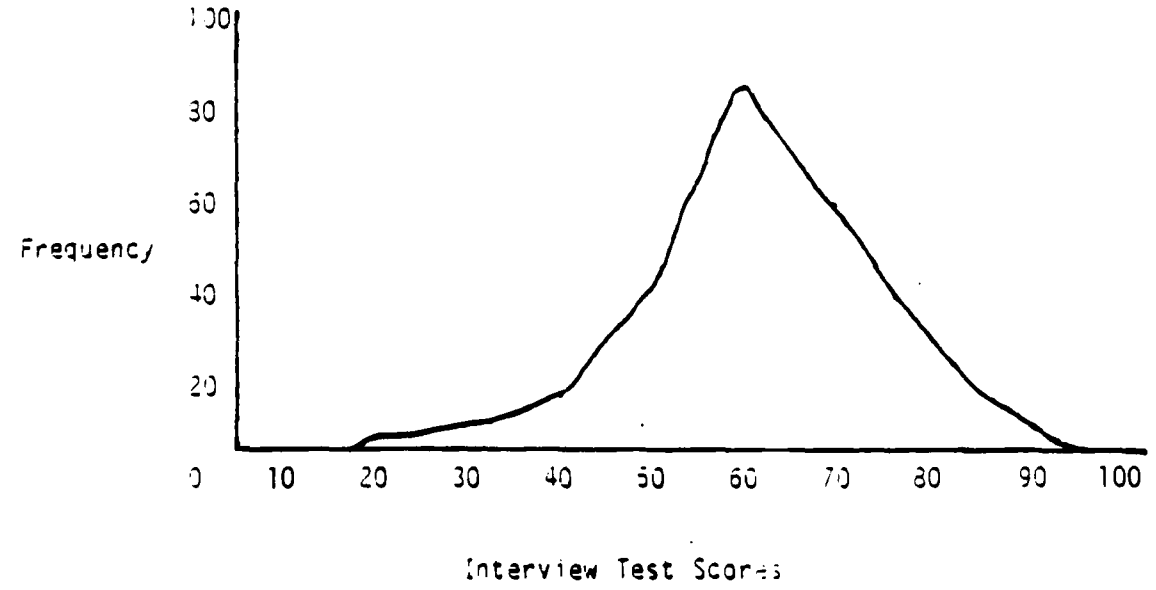

Figure 2. **Interview Work Sample Test Score Distribution.**

For a closer look at work sample test scores, Table 3 provides a task-by-task breakdown of means and standard deviations for both hands-on and interview items.

Table 3. Means and Standard Deviations for All Hands-On and Interview Work Sample Items

Item number	Hands-on		Interview	
	X	SD	X	SD
134	7.22	1.78	6.15	1.68
347	7.02	1.62	6.67	1.88
353	7.63	1.57	6.51	2.13
360	7.16	1.85	5.92	2.07
373	8.48	1.73	6.25	1.70
363	8.63	1.55		
301	6.22	1.91		
302	7.96	1.63		
349	5.55	2.22		
385	6.42	3.17		
346	7.04	2.27		
171	6.82	2.58		
351			6.70	2.20
359			6.38	2.15
396			7.00	2.01
387			8.20	1.60
319			4.71	1.98
239			7.26	2.66
247			7.95	2.16
238			6.54	2.12
208			7.03	1.80
325			5.24	1.84
328			7.00	2.44

Comparison of Hands-On and Interview Work Samples

Correlation Between Hands-On and Interview Tests. To assess how similarly the hands-on and interview tests rank order job incumbents, a bivariate correlational analysis was performed on summary scores from the 10 hands-on and 10 interview items. This analysis yielded a correlation of .57 between the hands-on and interview tests. As shown in Table 3, five hands-on and five interview items were also constructed for identical tasks. These five items yielded a correlation of .45 between the hands-on and interview work sample methods.

Correlation of Performance-Relevant Variables with Work Samples. In addition to assessing the correlation between hands-on and interview tests, an additional step was taken to compare hands-on and interview measures. This entailed examination of each work sample's relationship to other relevant variables (e.g., predictors, experience indices). A potential surrogate should be not only meaningfully related to the hands-on measure, but should also demonstrate similar patterns of relationships (as the hands-on) with other variables.

Correlations were calculated between the hands-on and interview work samples and 13 relevant variables as follows: six experience variables, collected via questionnaire, or gathered from personnel files; two motivation variables, consisting of job satisfaction (8 items) and supervisory support (2 items); final technical training school grade; and scores on the four aptitude composites from the Armed Services Vocational Aptitude Battery (mechanical, administrative, general, and electronics).

Table 4. Correlations Between Work Sample Tests and Performance-Relevant Variables

Performance-relevant variables	Work samples	
	Hands-on	Interview
Experience		
Time in Service	.19	.21
Time on Engine	.22	.20
Time in Unit	.18	.15
Number of Times Performed	.24	.27*
Last Time Performed	.19	.15
Task Experience	.21	.25
Motivation		
Job Satisfaction	.05	.06
Supervisory Support	.10	.03
Training		
School Grade	.22	.18
Aptitude		
Mechanical	.17	.22
Administrative	.09	.09
General	.11	.22
Electronics	.12	.25*

*Significant differences between hands-on and interview items are at the .05 level.

As shown in Table 4, a test of significance between hands-on and interview correlations produced only 2 significant differences across the 13 performance-relevant variables. This suggests similar patterns of relationships between the hands-on and work sample tests across a set of common variables.

CONCLUSION AND IMPLICATIONS

The major purpose of this paper was to explore the work sample testing philosophy and methodology being used by the Air Force in the Joint-Service Job Performance Measurement Project. In the development of objective measuring devices, a main requirement is freedom from biasing errors that arise from the performance situation and the method of measurement. Biasing factors in the performance situation (e.g., tasks performed, availability of tools and equipment) and the method of measurement (e.g., objectivity of work sample test administrators) are often found to systematically influence the quality of performance and performance measurement. Consequently, the developmental process, test administrator training, and data collection were discussed in some detail, and pertinent analyses presented to clarify the characteristics of the work sample tests. In addition, it was desirable to draw some conclusions about criterion equivalence between the hands-on and interview work samples; thus, a comparison of the two approaches was presented.

In general, both the hands-on and interview versions of Walk-Through Performance Testing held up well under close scrutiny. After thorough training, test administrators were able to rate reliably and accurately, over time, using both the hands-on and interview instruments. Both tests also showed moderate mean test scores and sufficient variability to suggest good discrimination among incumbents. Finally, two analyses examined hands-on and interview equivalence. A correlation of .57 provides additional strength to the interview testing approach

as a valid work sample methodology. In terms of relationships to a set of relevant variables, the interview showed a pattern similar to that of the hands-on test. Taken together, this information describes a solid program of work sample testing. Also, a first demonstration of interview testing was successful, suggesting this approach as a viable new work sample technology. Additional analyses are underway to examine the types of task characteristics that make some tasks more amenable to interview work sample testing. As data become available on three additional AFSs, a more precise understanding of hands-on and interview tests is possible.

III. PREDICTIVE EFFICIENCY OF THE ASVAB FOR THE AIR FORCE'S JOB PERFORMANCE MEASUREMENT SYSTEM

Terry L. Dickinson

Old Dominion University

Jerry W. Hedge
Lt Col Rodger D. Ballentine

Air Force Human Resources Laboratory

The Air Force Human Resources Laboratory (AFHRL) is currently developing a measurement system for obtaining job performance data. This Job Performance Measurement System (JPMS) will serve three interrelated purposes. First, the JPMS will provide operational managers of the Air Force's human resources program with criteria to evaluate program effectiveness. Second, the JPMS will provide Air Force research scientists with performance measures to use in research and development (R&D) projects. Finally, the JPMS will provide measures for assessing how well the Armed Services Vocational Aptitude Battery (ASVAB) predicts on-the-job performance. In this paper, we examine the underlying structure of the JPMS, and the predictive efficiency of the ASVAB for the JPMS measures.

BACKGROUND

Job performance is a complex concept. It consists of several dimensions that are predicted by many human attributes. The complexity of job performance has led researchers to advocate the use of multiple measures of job performance that are homogeneous in content and relatively independent of each other (Dunnette, 1963; Guion, 1976). This construct-oriented approach clarifies the conception of job performance and thereby enhances the understanding of predictors of job performance.

The Air Force's JPMS emphasizes an hierarchical classification of job performance, as well as multiple methods and sources for measurement (Kavanagh, Borman, Hedge, & Gould 1986). The broadest classification defines the components of job performance to reflect either (a) technical or (b) interpersonal aspects of work.

At the next level of hierarchy, job performance components are classified by dimensions. Each dimension still reflects technical or interpersonal performance; however, the concept of performance is enriched by subclassifying of technical and interpersonal performance into dimensions.

Many subclassifications of performance components into dimensions are possible, but their content usually emphasizes task, behavior, or trait information. The Air Force uses two subclasses: One emphasizes task-oriented and the other trait-oriented information.

The task-oriented dimensions are also subclassified. Each dimension is broken down into a set of interrelated tasks that reflect the content of the dimension. Furthermore, the tasks are broken down into task steps. These steps reflect the elemental or "go" versus "no-go" aspects of task performance.

The JPMS also emphasizes the use of multiple methods and sources for job performance information. Methods include testing, interviewing, or rating. Sources are the individuals who

provide the information, and they include peers, incumbents, supervisors, and experts. The Air Force uses testing and interviewing procedures to collect information using expert test administrators, as well as rating procedures to obtain information from incumbents, supervisors, and peers.

PURPOSE

The purpose of the paper was (a) to assess the structure of JPMS measures and, if necessary, revise the conception, in order (b) to evaluate the ASVAB's predictive efficiency in terms of these measures.

DATA COLLECTION

Participants

Performance data were collected on 255 first-term jet engine mechanics at 13 geographical locations in the United States early in 1985.

Job Performance Measures

Job performance information was obtained with testing, interviewing, and rating procedures. The interviewing and testing procedures are collectively referred to as Walk-Through Performance Testing (WTPT). The testing component is a traditional hands-on performance test that is administered by trained personnel. For example, a hands-on test for a jet engine mechanic requires the incumbent to install a starter on a jet engine. As the starter is installed, the test administrator uses a checklist to indicate whether each task step is performed correctly. The interviewing component is also administered systematically by trained personnel. It requires the incumbent to explain the step-by-step procedures that must be employed for successful task performance. The distinction between the hands-on testing and interviewing is clear. Hands-on testing emphasizes "can do" the task, while interviewing emphasizes "knows how to do." Hands-on data were obtained on 10 tasks, and interviewing data on 10 tasks. Five of the 20 tasks were common to both hands-on testing and interviewing. Separate scores were obtained for the unique hands-on and interviewing tasks as well as for the total 20 tasks. The WTPT scores provided an indication of technical proficiency.

Ratings were obtained for all the content levels of job performance from incumbents, supervisors, and peers. The task rating form provided the most specific rating data. Thirty tasks were rated with 5-point scales anchored with adjectives at each point.

The task-oriented dimensional rating form required incumbents, supervisors, and peers to rate technical proficiency on task-oriented dimensions. Potential dimensions were identified through factor analysis of occupational survey data. In a series of workshops with subject-matter experts, the dimension definitions and representative tasks were discussed, and 5-point rating scales were constructed for each dimension. Behavioral descriptions for each of the five points were developed using the behavioral summary statement approach advocated by Borman (1979). The dimensions were: (a) completion of forms; (b) remove/replace engine components; (c) inspect engine; (d) quality control; (e) shop maintenance; (f) preparation for storage and shipping; (g) flightline maintenance; and (h) troubleshooting. The dimension scores were averaged to obtain an indication of technical proficiency based on task-oriented dimensions.

The trait-oriented dimensional rating form was developed to be representative of all specialties in the Air Force. It focuses on traits that distinguish effective performers across

all jobs. The form was constructed by a group of "resource managers" in the Air Force who had managerial responsibilities for a large number of specialties. They were able in discussion to compare the performance requirements of several specialties and reach consensus on an inter-specialty perspective of performance. In addition, the managers developed 5-point rating scales that were anchored with behavioral summary statements. The dimensions were: (a) technical knowledge/skill, (b) initiative/effort, (c) knowledge of and adherence to regulations/orders, (d) integrity, (e) leadership, (f) military appearance, (g) self-development, and (h) self-control. For the trait-oriented rating form, the technical knowledge/skill dimension was retained as an indication of technical proficiency, while the remaining dimension scores were averaged to indicate interpersonal proficiency.

The global rating form was developed to measure technical and interpersonal proficiencies needed for successful performance. The two items were also developed in a workshop setting. The two types of proficiency were discussed and defined, and the behavioral summary approach was used to place specific behavioral descriptions on 5-point scales.

Other Measures

The ASVAB scores for the 255 participants were obtained from their personnel records. These scores included values for the 10 ASVAB subtests, the Armed Forces Qualification Test (AFQT), and the four Air Force composites (Mechanical, Administrative, General, and Electronics). The Mechanical composite is used to classify personnel to the Jet Engine Mechanic Specialty.

Two-hundred and three participants were tested with ASVAB Forms 8, 9, or 10 (which were operational between 1980 and 1984); 48 participants were tested with ASVAB Forms 5, 6, or 7 (operational between 1976 and 1980). One participant did not have the ASVAB form identified in his records, and three participants did not have ASVAB data in their records. The four participants with incomplete ASVAB data, and those tested with Forms 5, 6, and 7 were eliminated from the sample. Forms 8, 9, and 10 are calibrated to a 1980 reference population, and Forms 5, 6, and 7 are calibrated to a 1945 population. Thus, the two sets of forms should not be compared.

Training school grades were also available from the personnel records. Since training grades have frequently been used to describe the validity of ASVAB predictors, they served as a comparison criterion for the JPMS measures.

RESULTS

Structure of JPMS Measures

The hypothesized structure of the JPMS was evaluated using confirmatory factor analysis (Joreskog, 1971). In this approach to hypothesis testing, factor structure and factor intercorrelation matrices are specified to indicate the nature of the latent traits or factors that underlie the measures.

The hypothesized factor structure specified two general performance factors (i.e., technical and interpersonal proficiency) to underlie the JPMS measures. In addition, four factors were hypothesized as methods of measurement. These method factors were defined by the WTPT measures and the three sources who provided ratings (i.e., self, supervisor, and peer).

Table 5. Factor Structure of the Job Performance Measurement System

Variables	Factors				
	TECH	SUPER	SELF	PEER	INPERS
Hands-on	.973	.290	.182	.247	.012
Interview	.563	.184	.179	.182	-.169
T-self	.276	.288	.846	.269	-.017
T-supervisor	.282	.825	.356	.342	-.088
T-peer	.247	.414	.338	.838	-.057
D-self	.241	.340	.872	.295	-.081
D-supervisor	.368	.884	.452	.457	-.101
D-peer	.448	.437	.419	.855	-.135
A-self-TK	.248	.406	.806	.344	-.018
A-self-IP	-.078	.278	.577	.200	.570*
A-supervisor-TK	.299	.832	.398	.428	.029
A-supervisor-IP	.089	.823	.131	.454	.403
A-peer-TK	.352	.474	.292	.768	-.071
A-peer-IP	.079	.308	.091	.682	.433
G-self-TK	.258	.286	.747	.212	-.010
G-self-IP	.034	.232	.486	.209	.411*
G-supervisor-TK	.361	.803	.331	.457	-.049
G-supervisor-IP	.078	.678	.054	.394	.410
G-peer-TK	.352	.498	.353	.697	-.093
G-peer-IP	-.001	.311	.032	.617	.287

*Underlined loadings indicate marker variables for a factor. Abbreviations for factors are: TECH = proficiency of the technical aspects of work; SUPER = overall performance from supervisory point of view; SELF = overall performance from incumbent's point of view; PEER = overall performance from a peer point of view; and INPERS = overall performance on interpersonal factors from incumbent, supervisor, and peer points of view. Abbreviations for variables are T = task-level ratings; D = task-oriented dimensional-level ratings; A = Air Force-wide dimensional-level ratings; G = global-level ratings; TK = technical knowledge and skill; and IP = interpersonal aspects of work.

The factor intercorrelation matrix was hypothesized to contain correlations of zero between the performance and method factors. However, the matrix was hypothesized to contain nonzero correlations between the performance factors and between the method factors.

The hypothesized factor model provided a poor fit to the JPMS measures. Indeed, the maximum likelihood estimation procedures did not converge to a proper solution.

Next, an exploratory factor analysis procedure was used to discover the underlying structure of the JPMS measures (Joreskog, 1969). The results of this analysis are shown in Tables 5 and 6. As shown in Table 5, a five-factor solution was obtained that was similar to the hypothesized structure. The technical and interpersonal performance factors were obtained, as well as three factors for the three sources of ratings. However, a separate factor for the WTPT measures did not appear (i.e., hands-on and interview). These measures along with task, task-related dimension, and global technical ratings from the three sources defined the technical performance factor.

As shown in Table 6, the factors had low to moderate correlations. The largest correlation occurred between the supervisor and peer factors (i.e., .474), and it agrees with previous research (e.g., Klimoski & London, 1974; Lawler, 1967).

Table 6. Correlations Among the Factors of the Job Performance Measurement System

Factors	TECH	SUPER	SELF	PEER	INPERS
TECH	1.000				
SUPER	.268	1.000			
SELF	.234	.319	1.000		
PEER	.248	.474	.254	1.000	
INPERS	-.206	.107	-.001	.112	1.000

Note. Abbreviations are defined in Table 1.

Predictive Efficiency of ASVAB for JPMS

The ability of the ASVAB to predict the JPMS measures was evaluated using correlations that were and were not corrected for range restriction. As noted by advisory groups to the Joint-Service Job Performance Measurement Project, the corrected correlations provide a common basis for interpretation. The base group used for correction was the 1980 Youth Population (Department of Defense, 1982), and the Pearson-Lawley procedure was used for multivariate correction on the 10 subtests of the ASVAB (Mifflin & Verma, 1977).

Table 7 contains subtest intercorrelations for the 1980 Youth Population (upper triangle) and for the sample of jet engine mechanics (lower triangle). A comparison of the correlations indicates large discrepancies between some correlations. For example, the correlation between general science (GS) and arithmetic reasoning (AR) for the Youth Population is .722, while for the mechanics it is .221. Such discrepancies indicate that the results for corrected correlations should be interpreted cautiously.

Table 7. Intercorrelations Among the ASVAB Subtests for the 1980 Youth Population (Upper Triangle) and for the Sample of Jet Engine Mechanics (Lower Triangle)

	ASVAB Subtests									
	GS	AR	WK	PC	NO	CS	AS	MK	MC	EI
GS	1.000	.722	.802	.690	.525	.452	.637	.695	.696	.760
AR	.221	1.000	.709	.671	.626	.515	.533	.828	.685	.658
WK	.622	.296	1.000	.803	.618	.552	.529	.671	.595	.684
PC	.369	.281	.516	1.000	.608	.561	.423	.637	.522	.572
NO	.063	.450	.145	.256	1.000	.701	.307	.617	.409	.422
CS	.025	.267	.033	.104	.497	1.000	.225	.520	.336	.341
AS	.342	.217	.263	.207	-.019	-.022	1.000	.415	.741	.746
MK	.285	.695	.352	.391	.465	.312	.092	1.000	.601	.585
MC	.326	.457	.338	.226	.134	.139	.422	.416	1.000	.744
EI	.402	.233	.385	.291	-.002	.082	.530	.235	.502	1.000

Note. Abbreviations for the subtests are GS = General science; AR = Arithmetic reasoning; WK = Word knowledge; PC = Paragraph comprehension; NO = Numerical operations; CS = Coding speed; AS = Auto and shop information; MK = Mathematics knowledge; MC = Mechanical comprehension; and EI = Electronics information. All sample correlations greater in absolute size than .138 are significant at $p < .05$.

The correlations between ASVAB predictors and JPMS measures are reported in Tables 8 and 9. Scores for the JPMS factors were obtained by a weighted average of the marker variable scores. All the marker variables had a weight of 1.0, except the hands-on measure. It was weighted 2.0 because of its greater importance in defining the factor of technical proficiency.

The ASVAB modestly predicted the JPMS measures. The corrected and uncorrected correlations suggest that the Mechanical and Electronics composites, as well as the 10 subtests, are most predictive of the JPMS measures. Furthermore, the magnitude of the Electronics and subtest correlations relative to those of the Mechanical composite suggest that the ASVAB has additional predictive power that could be used to classify personnel to the Jet Engine Mechanic specialty.

Table 8. Correlations Between ASVAB Predictors and JPMS Measures (Uncorrected for Restriction)

JPMS measures	ASVAB predictors					
	AFQT	M	A	G	E	Subtests[a]
TECH	.163	.214*	.099	.170	.192*	.338**
SUPER	.080	.111	.074	.085	.130	.238
SELF	.034	.074	.043	.000	.116	.319*
PEER	.124	.090	.124	.114	.112	.201
INPERS	.107	.043	.104	.077	.068	.170
TOTWTPT	.150	.207*	.096	.158	.181*	.334**
Hands-on	.120	.170	.086	.110	.122	.311*
Interview	.187*	.224*	.091	.219*	.251**	.328**
TGRD	.477**	.445**	.376**	.490**	.530**	.581**

Note. Abbreviations for composites are AFQT = Armed Forces Qualification Test; M = Mechanical; A = Administrative; G = General; and E = Electronics. The abbreviation TOTWTPT is for the total score obtained with Walk-Through Performance Testing and TGRD is the grade received in technical training. See Tables 1 and 3 for the remaining abbreviations.

[a]The values reported for the subtests are multiple correlations.

*$p < .05$.

**$p < .01$.

The ASVAB was a somewhat better predictor of the interviewing measure than of the hands-on measure. A probable explanation for this finding is a common requirement of verbal ability for ASVAB predictors and the interviewing measure. The interview requires the incumbent to "show and tell" how to do the task, whereas the hands-on measure requires the incumbent "to do" the task.

The ASVAB did better in predicting training school grades than it did in predicting JPMS measures. Historically, the ASVAB has been revised on the basis of its relationship to training school grades; so, this finding is not surprising. However, it is important to assess whether the ASVAB has any predictive efficiency that is unique to measures of on-the-job performance.

Finally, a comparison of uncorrected and corrected correlations in Tables 8 and 9 indicates that several correlations were reduced in magnitude upon correction. These reductions were apparently due to greater standard deviations for some of the ASVAB subtests in the jet engine mechanic data compared to Youth Population data.

Table 9. Correlations Between ASVAB Predictors and JPMS Measures (Corrected for Restriction)

JPMS Measures	ASVAB predictors					
	AFQT	M	A	G	E	Subtests[a]
TECH	.219	.244*	.159	.220	.237*	.306**
SUPER	.175	.152	.215	.166	.188	.289
SELF	-.134	-.051	-.191	-.153	-.089	.357*
PEER	.324	.232	.350	.300	.292	.368
INPERS	.211	.135	.190	.180	.173	.235
TOTWTPT	.232	.245*	.193	.231	.244*	.307**
Hands-on	.204	.213	.159	.188	.198	.292*
Interview	.180*	.224*	.111	.209*	.232**	.276*
TGRD	.621**	.543**	.582**	.620**	.627**	.671**

Note. Abbreviations are defined in earlier tables. Statistical significance reported for uncorrected correlations in Table 4 is inferred for correlations in this table.

[a]The values reported for the subtests are multiple correlations.

*$p < .05$.

**$p < .01$.

Several sets of multivariate regression analyses, as reported in Tables 10 and 11, were performed to assess the unique predictive efficiency of the ASVAB subtests. The WTPT measures were analyzed separately and as a composite. In addition, two orders of entry were specified for Roy-Bargman step-down tests (Bock, 1975). One order addressed the question: Does the ASVAB predict performance in training that is uniquely different from on-the-job performance? Sets 1 and 2 addressed this question. The second order addressed: Does the ASVAB predict job performance uniquely beyond that of training performance?

Table 10. Summary of Roy-Bargman Step-Down Tests (Uncorrected for Restriction)

Analysis		df		Mean squares		
Set	Order	HYP	MSE	HYP	MSE	F-ratio
1	TOTWTPT	10	178	437.88	222.81	1.96*
	TGRD	10	177	258.50	29.33	8.81**
2	Hands-on	10	178	196.42	104.90	1.87
	Interview	10	177	104.51	107.07	.98
	TGRD	10	176	259.73	29.22	8.88**
3	TGRD	10	178	283.80	30.06	9.44**
	TOTWTPT	10	177	333.98	217.45	1.53
4	TGRD	10	178	283.80	30.06	9.44**
	Hands-on	10	177	177.38	101.42	1.75
	Interview	10	176	82.70	107.67	.77

Note. Abbreviations are HYP = Hypothesis and MSE = Mean square error. The remaining abbreviations are defined in earlier tables.

*$p < .05$.

**$p < .01$.

As shown by the step-down tests in sets 1 and 2, the training performance predicted by ASVAB is uniquely different from on-the-job performance. The step-down tests shown for sets 3 and 4 suggest that the ASVAB does not predict unique on-the-job performance beyond that accounted for in training grades. These results occurred for both corrected and uncorrected correlations.

Table 11. Summary of Roy-Bargman Step-Down Tests (Corrected for Restriction)

Analysis		df		Mean squares		
Set	Order	HYP	MSE	HYP	MSE	F-ratio
1	TOTWTPT	10	178	422.51	227.72	1.86*
	TGRD	10	177	366.23	28.47	12.86**
2	Hands-on	10	178	174.98	105.75	1.65
	Interview	10	177	84.84	105.18	.81
	TGRD	10	176	377.10	28.17	13.38**
3	TGRD	10	178	421.91	29.00	14.55**
	TOTWTPT	10	177	194.44	223.60	.87
4	TGRD	10	178	421.91	29.00	14.55**
	Hands-on	10	177	120.64	105.78	1.14
	Interview	10	176	59.17	102.78	.58

Note. The abbreviations are defined in earlier tables. Statistical significance reported for uncorrected tests in Table 10 is inferred for tests in this table.

*$p < .05$.

**$p < .01$.

DISCUSSION

This paper has presented an investigation into the structure of the JPMS. The results were encouraging, in that five of the six factors obtained were previously hypothesized to describe the JPMS. The WTPT factor did not emerge, but it was part of the hypothesized factor of technical proficiency. This interpretation is supported by the loadings of the rating measures on the technical proficiency factor. Substantial loadings were obtained for all the rating measures that tapped technical proficiency (factors TECH, SUPER, SELF, and PEER in Table 1), while near-zero loadings were obtained for the rating measures that tapped interpersonal proficiency (factor INPERS in Table 1).

The interpersonal factor also appeared as hypothesized. Although the factor loadings of marker variables were smaller in magnitude than those for technical proficiency, a common interpersonal proficiency factor was found in ratings by incumbents, supervisors, and peers. Perhaps, this factor will appear stronger in the application of the JPMS to Air Force specialties that are service-oriented (e.g., Personnel technician).

The meaning of the source-of-rating factors remains a topic for future research. The loadings of the self, supervisory, and peer factors suggest that both technical and interpersonal aspects of the work are being tapped. Of course, the factors may indicate nothing more than unique sources of bias. If so, the general lack of predictive efficiency of the ASVAB for these factors is understandable.

This paper also investigated the ability of the ASVAB to predict the JPMS measure. In general, the results indicated modest success in predicting on-the-job performance. Only the WTPT measures appeared to be predicted by the ASVAB.

Of the ASVAB composites, the Mechanical and Electronics appeared to be most valid. This occurred for both uncorrected and corrected correlations. This lent some confidence to the use of corrected correlations. However, until JPMS results from other specialties are available, the interpretation of corrected correlations must remain cautious.

In conclusion, the results of this paper indicate that (a) JPMS has a meaningful structure which is quite similar to the hypothesized structure, and (b) ASVAB predictors are only modestly predictive of the on-the-job performance of jet engine mechanics.

IV. AIR FORCE JOB PERFORMANCE MEASUREMENT TECHNOLOGY APPLIED TO TRAINING

Jack L. Blackhurst
Rodger D. Ballentine
Martin W. Pellum

Air Force Human Resources Laboratory

OVERVIEW

All Air Force training programs are evaluated in some way, whether it be asking trainees their opinions of the training, or assessing the effect of the training on some work-related variables. Since most training is aimed at affecting job performance in some way, obtaining indices of job performance is critical to both the initial design of the training program, and the subsequent evaluation of the program's success. Resident technical training schools rely on written and hands-on tests to determine how well the training objectives are being met. These schools also depend on inputs from supervisors of recent graduates to ensure the curricula are preparing the students to effectively meet job demands. Supervisors and inspection teams use performance appraisals in making decisions about trainee proficiency and training program soundness. This paper addresses how performance measurement is currently utilized in resident technical training and in on-the-job training, and how advances in job performance measurement could aid in these endeavors. (Although the technical training comments in this paper focus on initial training, the concepts apply equally well to later phases of resident technical training.)

PERFORMANCE MEASUREMENT IN TRAINING TODAY

Initial Resident Technical Training

Initial resident technical training closely parallels the educational environment of civilian vocational schools. Courses are structured in their length, presentation methods, and evaluation processes. Students learn in both classroom and laboratory settings. The training curriculum is derived directly from the Specialty Training Standard (STS) using task analysis and Instructional Systems Development practices (AFM 50-2; AFP 50-58; DeVries, Eschenbrenner, & Ruck, 1980) and is intended to prepare students for a variety of potential jobs within a specialty (ATCR 52-3). Thus, the training is typically aimed at producing graduates who are semi-skilled or "partially proficient" on many tasks, rather than skilled or "proficient" on a few tasks. This fact introduces some unique considerations when applying job performance measurement technologies, which will be discussed later in this paper.

In resident technical training, performance appraisal plays a key role in evaluating success of both the students and the training program. The primary purpose of student performance measurement within the resident school is to ensure the training objectives are being met so the student may proceed through the course or graduate. However, such information is also used to evaluate the quality of the training itself.

Student accomplishment of each course objective is formally assessed via written and/or hands-on performance tests. Such checks typically involve demonstrating skills and knowledge required for task performance to the "partially proficient" level. For example, with hands-on performance tests, students may receive assistance from instructors or they may perform only the less difficult parts of the task. When tasks require a team effort, students may not have the opportunity to perform all aspects of a task, and therefore, evaluation reflects their team participation more than their individual performance.

One informal assessment technique used throughout resident training is the verbal quiz. During class discussion or task performance, students explain where they would go for information, what they are doing, why they are doing it, etc. This informal evaluation seems to be a powerful measure of student understanding of the underlying systems concepts.

Evaluation of resident technical training programs also derives from appraisals of graduate performance on the job through supervisor questionnaires, training quality reports (TQRs), and training center field visits (ATCR 52-12). About 3 months after completing technical school, graduates and their supervisors complete rating forms covering how well the training program prepared the graduates for duty, using the STS task listing as a guide. The TQR program enables anyone to express an observation concerning training program results. Usually this is a perceived training deficiency in graduates, either where STS standards are not being met or the training is inadequate given on-the-job requirements. Finally, field visits are conducted by a training evaluation team in which recent graduates are interviewed and often asked to perform certain tasks taught in the resident course. All of this external performance information is fed back to the resident training managers for consideration and action, if necessary. It is also combined with other (internal) training information (test scores, washback rates, etc.) and compiled into a training evaluation report (AFR 50-38).

On-the-Job Training

The on-the-job training (OJT) component is considerably different from resident technical training. OJT is "dual-channeled" in that individuals receive training on their entire specialty through career development courses (CDCs), and training on their particular jobs through interactions with supervisors/trainers. The training occurs in the operational environment where newly assigned trainees learn a job and contribute to the mission at the same time. This places different constraints on OJT than those experienced in initial technical training; as job pressures increase, training may suffer. Though guidance on how to conduct OJT is outlined in several different places (e.g., AFR 50-23, AFM 66-1), OJT remains a flexible, albeit unstructured process. The STS serves primarily as a guide for the level of proficiency required for task certification, while the supervisor has discretion over what is trained. (In some cases, major commands may require that certain tasks be included in this training by issuing command-level job performance guides.)

Typically, OJT is administered in a one-on-one fashion, with an expert training a novice. Generally, there are limited instructional materials, and training practices may vary greatly from one supervisor/trainer to another. Performance evaluation practices may vary greatly as well. Prior to certifying a trainee on a task, the supervisor must ensure the trainee is able to perform at the given level of proficiency. The supervisor may evaluate trainee job performance any number of ways, from hands-on demonstrations to inspections of end products. CDC performance is measured by written volume review exercises and course examinations.

Performance appraisals are also used by base and major command inspection teams. Individuals are selected to perform certain tasks, and their performance is evaluated against the standards found in the applicable regulations and technical orders. These individual evaluations could be used as a check on the training system as a whole.

APPLYING PERFORMANCE MEASUREMENT TECHNOLOGY TO TRAINING

Background

In 1983, the Air Training Command formally expressed the need for research linking resident technical training to individual job performance. Two separate research efforts exploring the

potential integration of performance measurement technologies in the technical training evaluation process have been completed since that time (Banks, 1987; Driskill, Mitchell, & Ballentine, 1985). As a result of these two studies, there exists a clearer notion of how and where the job performance measurement process and information could benefit the technical training community. Research on performance measurement in OJT arose from a different set of circumstances. Following an AF-wide inspection of the entire OJT system in 1977, work was initiated to improve all aspects of OJT through the application of state-of-the-art computer and automation technology (Carson, Chambers, & Gosc, 1984). Within this paper, comments on how and where performance measurement could benefit OJT will be confined to applications within the Advanced On-the-job Training System (AOTS).

Initial Resident Technical Training

The resident technical training process can be divided, for discussion purposes, into three phases. The nature of the performance information gathered and its use may differ depending on which phase is targeted. Therefore, each phase will be treated separately, even though there are obvious overlaps between them.

Phase I: Pre-Training

The processes that occur within this phase all relate to development of the training program. This phase encompasses the identification of tasks performed by the target group (usually first-term airmen), and the translation of this information into training objectives, and ultimately lesson plans. In essence, it includes anything that occurs prior to the active training of individuals.

Two areas in this phase could benefit from measures of job performance. First, the validity of the Instructional Systems Development (ISD) model could be determined by empirically linking the developed training curriculum to job performance. Though this feedback loop currently exists, the methods used (i.e., surveys and interviews) have not been compared to more rigorous methods such as performance tests. (See the Phase III discussion below.) Thus, one cannot be sure of the quality of the information that is currently gathered. Second, the behaviors associated with given levels of proficiency may or may not be the critical ones needed upon entry on the job. Since not all aspects of task performance can be covered in resident training, information is needed that would enable trainers to more precisely define the course content, in a manner that best meets the job requirements, supervisor expectations, and training constraints. Job performance measures could give this entry-level task proficiency information.

Phase II: Training

This phase reflects the action of instructing, material presentation, involvement of the trainees, and the methods used to assess trainee progress. Since much of the success of a training program rests with the ability of the instructor to convey the material, research on methods for "training the trainer" and evaluating instructor performance appears warranted. Performance measurement development methods and formats from the emerging performance technology could assist in this area. Student measurement is another area that could benefit from this technology. Though the present internal training evaluation methods appear to work, their validity and reliability are rarely assessed. Additional or different methods might yield greater information or increase the validity of information collected, leading to better administrative decisions.

Phase III: Post-Training

Determining whether the training program improved subsequent job performance is the key function of this phase. It forms the external feedback loop needed to validate the training process begun in Phase I. Several different approaches have been taken to empirically relate what is trained to what needs to be trained. One could use surveys or subject-matter expert (SME) judgments. Pennell, Harris, and Schwille, (1976) developed and demonstrated a factor-analysis-based methodology for using the existing supervisor surveys of recent resident technical training graduates' performance. Based on supervisor ratings of performance, they identified areas that were potentially being overtrained or undertrained. Ford and Wroten (1984) achieved the same basic end through the calculation of content validity ratios (CVRs) using SME inputs. They used these CVRs to match training needs to current training emphases at the training category and subcategory levels.

Very little work exists that relates hands-on measures of job performance to training program content or methods. For this reason, training course developers often lack the information needed to make decisions on how the training curriculum can be structured to provide the greatest transfer of learning to job. One could obtain measures of productivity or effectiveness, such as maintenance downtime or sortie flying rates, and relate these to the training. These "production functions" would enable one to estimate the costs and benefits associated with particular training program designs (Solomon, 1986). However, these measures are often difficult to obtain, are affected by numerous other factors besides training, and have been shown to not always be reliable (Gibson & Orlansky, 1986).

Another way to obtain the needed job performance information would be to develop tests of tasks individuals perform in their jobs. The Air Force's Walk-Through Performance Testing (WTPT) technology (Hedge & Teachout, 1986) has been used to obtain such measures of job performance for selection validation purposes (Dickinson, Hedge, & Ballentine, 1987). This same information has also been related to resident technical training achievement (Hedge, Ballentine, & Gould, 1985). What is yet to be developed, however, is a WTPT-like system geared specifically toward identifying training needs.

On-the-Job Training

It is conceivable that measures could be developed that could evaluate the training conducted prior to an individual reaching his/her job and identify additional OJT training needs. If designed for and administered to recent resident technical training graduates, the results of such measures could be fed back to the training center and used by the supervisor to determine the point at which to start the person in OJT. If the measures reflected tasks learned primarily in OJT, the results could be used by an individual's new supervisor as a diagnostic tool in identifying the extent of training necessary to make the person qualified for a position held. Both of these concepts fit within the Advanced On-the-job Training System (AOTS) scenario discussed below.

The Advanced On-the-Job Training System

The AOTS will provide the capability of real-time management of all aspects of OJT to include automatic record-keeping and updates, as well as scheduling and tracking trainee progress. In addition, it will provide specialty training requirements at multiple levels (e.g., by position, unit, or specialty). The system uses a computer management system with both training and evaluation components (see Figure 3).

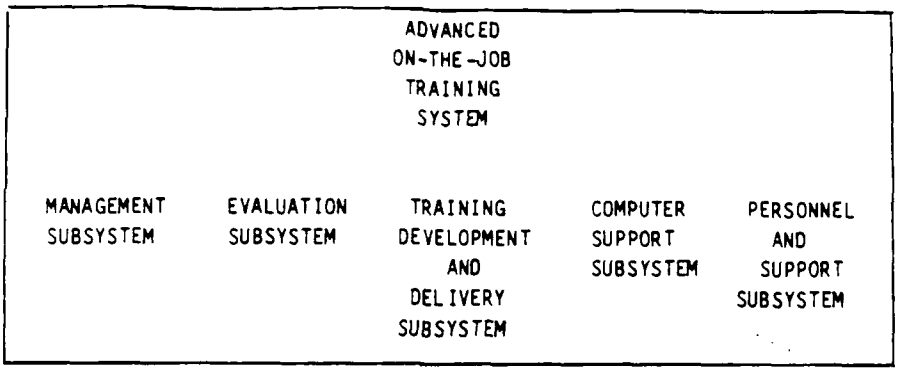

Figure 3. AOTS Subsystems.

The training and evaluation components will provide the Air Force with the capability of on-line training and evaluation. SMEs will identify the task training requirements and actually build the training and evaluation materials for supervisors--making it a real "blue-suit" system. Training requirements can then be matched to a particular position, rather than to an entire specialty as is currently done. The results of the evaluations will be updated automatically or through optical scan devices. The system is designed to be available to the supervisor and trainee in the work center, where OJT is accomplished today. The system should allow training to occur without the need for a supervisor to be present for every task, since computer-based training will be used for some tasks. For those critical tasks that need one-on-one training, the supervisor should be able to spend more quality training time with the trainee.

V. INTER-SERVICE TECHNOLOGY TRANSFER: PROMISE AND PAYOFF

Robert E. Duncan
Dorian A. Hodge
Jack L. Blackhurst

Air Force Human Resources Laboratory

As the lead Service for job performance measurement (JPM) technology transfer, the Air Force has initiated a variety of efforts to investigate the feasibility not only of sharing previously developed measures but also of sharing developmental approaches. In the past, the major rationale behind technology transfer has been cost savings (Alba, 1986, p. 1). Today, that rationale remains. However, as we look toward the future of performance measurement technology, greater emphasis needs to be given to planning and executing a plan that expands the rationale for the transfer of performance measurement technology beyond cost savings/cost avoidance. The transfer of previously developed instruments has great potential, but the greater promise and payoff lay in transferring Service-specific measurement approaches to provide optimum validation of Armed Service Vocational Aptitude Battery (ASVAB) scores.

This paper will examine: (a) the scope of technology transfer; (b) past, present, and future transfer activities; and (c) the promise of future JPM technology transfer and its potential payoff to the Department of Defense (DoD) and private industry.

BACKGROUND

The DoD, in response to a Congressional mandate, is coordinating an effort among the Services to develop job performance measures for the validation of ASVAB. This effort is being directed by the Joint-Service Job Performance Measurement Working Group (JSJPMWG) composed of Service representatives and a representative from the National Academy of Sciences, the organization responsible for providing technical oversight and advice. Early in the effort, the JSJPMWG determined a need existed for ensuring JPM technology was shared (transferred) among the Services to the maximum extent possible. As a result, the Air Force was selected as the lead Service for technology transfer. The transfer efforts, initiated by the Air Force in 1985, are ongoing and will be more thoroughly discussed later in this paper.

THE SCOPE OF TECHNOLOGY TRANSFER

JPM technology transfer can be described as a two-pronged effort. We call the first component "technology transfer," where technology refers to the procedures the Services use in developing their respective hands-on and surrogate measures. In some Services these procedures are well documented, while in others the documentation is currently being prepared or is entirely absent. Although these last two conditions make technology transfer more difficult, it is important to pursue the transfer of Service-specific approaches to determine where modifications need to be made. The approach being followed to transfer Service-specific technology to other Services involves the active and cooperative interaction between both the donating and receiving Services.

Through modification of task sampling and analysis techniques (if required), the receiving Service follows the specific procedures used by the donating Service to develop the latter's surrogate measures (as transfer, at this time, is centered around surrogates). Such a procedure ensures that a receiving Service is developing a surrogate measure that would exactly parallel the surrogate developed by the donating Service. Technology transfer, as described here, will

yield the greatest long-term payoff--a series of JPM technologies usable not only by the Services but also by the private sector. As a receiver of technology, the Air Force is attempting to document the development of another Service's surrogate in an ongoing transfer effort. Such action is required to design a truly uniform Joint-Service JPM technology.

The second component of transfer involves what we call "transfer-in-kind." Transfer-in-kind is simply the application, with minor modifications of nomenclature, of the performance measures developed for a specific specialty of the donating Service to a very similar specialty of the receiving Service. The major benefit of this type of transfer is short-term cost avoidance (i.e., avoiding the costs associated with the full development of performance measures in similar specialties). Past transfer efforts, to be described later, have focused on this type of transfer in order to: (a) demonstrate the feasibility of inter-Service JPM technology transfer, and (b) save developmental costs in a time of austere funding. While the feasibility question has been resolved in the positive, the lack of procedural documentation severely impacts long-term use of the transferred technologies.

EXAMPLES OF TECHNOLOGY TRANSFER

Transfer of Jet Engine Mechanic Instruments to the Navy

Background

In 1985, an effort was begun to transfer the Air Force jet engine mechanic (JEM) Job Performance Measurement System (JPMS) to the Navy. This transfer effort is thoroughly documented in Alba (1986), Baker and Blackhurst (1986), and Blackhurst and Baker (1985). However, a brief review of purpose, approach, and cost/benefit analyses may be needed. The JEM technology transfer was a "transfer-in-kind." By providing Air Force-developed JPMS instruments to the Navy/Marine Corps for use with their J-79 JEMs, the Navy was able to speed up the cradle-to-grave (from job domain definition to data analysis) process.

Approach

The Air Force provided the Navy with a completed JPMS for J-79 JEMs. The Air Force JPMS is composed of a Walk-Through Performance Test (WTPT), four types of rating forms, four questionnaires seeking attitudinal and motivational information, and instructions for the administration of all instruments. The WTPT contains both hands-on and interview items. Hands-on items are designed to permit a trained administrator to observe the performance of a job incumbent on specific tasks or parts of tasks which are part of a normal first-termer's job. Interview tasks require the incumbent to describe how a task should be accomplished (in a show-and-tell approach).

A representative set of items are present in both hands-on and interview modes to permit comparison between the hands-on benchmark and the interview surrogate. Rating forms include: (a) a global rating of both technical and interpersonal skills, (b) a rating form which examines military-related (as separate from job-related) performance (e.g., leadership, integrity), (c) a rating of all tasks reflected in the WTPT at the task level, and (d) a rating of tasks grouped into logical dimensions (e.g., removing and replacing components, troubleshooting). These different rating forms are administered to a job incumbent, the incumbent's coworkers, and the incumbent's supervisor, after all raters have undergone an extensive 4-hour training session. The JPMS is used to validate the ASVAB selector aptitude index (mechanical, administrative, general, and electronics) for a specialty. The Air Force-developed JPMS was evaluated by Navy, Marine Corps, and contractor personnel to ensure that the measurement system adequately measured

Navy/Marine Corps J-79 JEM performance. After review, minor revisions were made to adapt the instruments to reflect terminology commonly used in the Navy/Marine Corps. Additionally, two tasks not performed by Navy/Marine Corps JEMs were removed, and tasks frequently performed by these personnel were substituted. Due to an insufficient number of Navy JEMs, the revised instruments were administered to first-term Marine J-79 mechanics. Results of these data are currently being compiled. In addition to the revised Air Force instruments, a Navy job knowledge test was developed and administered to the Marine J-79 incumbents.

Cost Avoidance as a Result of Transfer-in-Kind

Baker and Blackhurst (1986) compared the costs associated with development of a JEM JPMS in the Air Force to costs incurred as a result of a transfer-in-kind to the Navy/Marine Corps. Table 12, reproduced from Baker and Blackhurst (1986), illustrates the results of this comparison. A complete verification of these costs was restricted due to the unavailability of necessary reference documents. Regardless, it can be seen that the transfer-in-kind to the Navy yielded an estimated 400% cost avoidance. However, the size of future cost avoidance figures may be different from that obtained by Baker and Blackhurst.

Table 12. Jet Engine Mechanics JPMS Transfer Cost Comparison

	Air Force Development		Transfer to Navy	
	Man-hours	Dollars	Man-hours	Dollars
Active Duty	3,797	$ 27,000	260	$ 5,200
Civilian	2,509	80,300	102	2,300
Contractor	4,164	143,000	1,860	46,000
Total	10,470	$250,300	2,222	$53,500

Conclusions

Although the final results of this transfer-in-kind effort are not yet in, the work completed so far indicates that this method of transfer is not only feasible but also accelerates cross-Service developmental efforts and avoids the costs associated with simultaneous development of separate JPM instruments for similar Service specialties. Transfer-in-kind, however, should not be limited to those specifically interested in JPM research across the Services. Each Service may have needs in their respective operational communities for instruments developed in the Joint-Service JPM project. Private industry may also need these same instruments but may lack the expertise for revision/development of performance measures. The Services should make an effort to provide industry with completed instruments to determine whether the measurement approach is usable outside a military framework.

Army Job Knowledge Test Transfer

Background

Upon recommendation of the National Academy of Sciences (NAS), the Air Force is developing Army Job Knowledge Tests for four Air Force specialties. This effort represents the first type of transfer--transfer of technology--and, as such, requires that great care be taken to ensure Air Force test developers understand the concepts which underlie the procedures used by Army

researchers to develop Army Job Knowledge Tests. This effort is facilitated by ongoing Air Force efforts to develop JPMSs for four Air Force specialties. Development of Job Knowledge Tests will occur simultaneously with Air Force JPMS development for these four specialties.

The approach being taken to develop Job Knowledge Tests includes consultation with Army representatives throughout development. Job Knowledge Test developers working for the Air Force will initially review documents provided by the Army that outline test content selection and test development procedures. In addition, test developers will review the test content resulting from Air Force JPMS development. After background information has been reviewed, test developers will meet with Army researchers to discuss all aspects of Job Knowledge Test development including task sampling procedures, test domain definition, test development, and methodological requirements. Once the instruments have been developed, test developers will again confer with Army researchers, and based upon their input, make any revisions necessary to ensure the accuracy of the Job Knowledge Tests as job performance measures. After development of the Job Knowledge Tests for the four designated Air Force specialties has been completed, the data collection and data analysis phases will begin. Approximately 250 subjects in each of the four career fields will be administered the Job Knowledge Tests along with the Air Force JPMS and a battery of additional predictors which include the Air Force Apprentice Knowledge Test for each career field and the Space Perception Test. A counterbalancing technique for test administration will be developed to guard against bias resulting from test administration order. Data analysis will examine, among other things, test reliability, validity, and the fidelity of each surrogate job performance measure when compared to the hands-on benchmark. The JSJPMWG will assist in developing a data analysis strategy that thoroughly assesses the potential utility and the most advantageous applications of surrogate job performance measures in future job performance measurement research. JSJPMWG assistance during both test development and data analysis will ensure a forward-looking perspective for technology transfer among the Services.

Anticipated Results

The most important result from this effort will be a demonstration of technology transfer. Technology transfer has not yet been attempted by the Services and may prove more difficult than transfer-in-kind because it will involve transfer of abstract concepts (procedures, methods, and approaches), rather than concrete information (instruments already developed). Even with the Army consulting throughout the development process, it is important to consider whether it is possible to transfer technology from one Service to another and still maintain the uniqueness of the donating Service's technology.

Several things may impact this transfer. First, job performance measurement technology can be used within each Service for additional payoffs, although each of the Services has as its overall objective the validation of enlistment standards under the Congressional mandate. For example, the Air Force plans to use job performance measurement technology extensively in the area of training. Additional within-Service applications of job performance measurement technology may overlap among the Services, but at the same time stem from Service-unique research requirements.

Second, inter-Service differences in test content definition, task sampling procedures, etc. may impact technology transfer, since each Service is developing both hands-on measures and Service-unique surrogate measures. However, we are looking specifically at methods of technology transfer which will transcend these differences.

If successful, this transfer may facilitate and promote future job performance measurement technology transfers among the Services. An important product resulting from transfer of the Army technology to the Air Force will be a report documenting procedures followed, problems

This paper also investigated the ability of the ASVAB to predict the JPMS measure. In general, the results indicated modest success in predicting on-the-job performance. Only the WTPT measures appeared to be predicted by the ASVAB.

Of the ASVAB composites, the Mechanical and Electronics appeared to be most valid. This occurred for both uncorrected and corrected correlations. This lent some confidence to the use of corrected correlations. However, until JPMS results from other specialties are available, the interpretation of corrected correlations must remain cautious.

In conclusion, the results of this paper indicate that (a) JPMS has a meaningful structure which is quite similar to the hypothesized structure, and (b) ASVAB predictors are only modestly predictive of the on-the-job performance of jet engine mechanics.

Training Evaluation

Job performance measurement definitely has an application to training evaluation. The Air Force currently uses a variety of performance measures to assess the effectiveness of its training Air Force-wide. These measures vary from hands-on measurement to surveys or interviews with students and trainees. The most common type of performance measurement is a paper-and-pencil knowledge test. However, in field settings, the actual hands-on evaluation is more commonly used. Both of these techniques are widely used in the DoD performance measurement research program, and results of their findings should enhance training evaluation policy and procedures in the Air Force. Two Air Force training research projects, the Advanced Training System (for technical schools) and the Advanced On-the-Job Training System (for on-the-job training), will have performance evaluation components which will use a variety of performance techniques. Procedures and lessons learned from the DoD effort can and will be applied to the training evaluation area.

Transfer to the Private Sector

One of the greatest payoffs of inter-Service technology transfer is providing both the technology and developed instruments (transfer-in-kind) to the private sector. Based on the successes achieved so far and planned future endeavors, there should be a substantial pool of inter-Service JPM technology which, if shared with industry, would significantly aid the DoD in establishing performance requirements for its contractors. This should enhance productivity and product quality throughout weapon system acquisition activities. In the future, the Air Force will be developing a JPMS for firefighters and a medical specialty. Not only will the other Services benefit from these developmental efforts but DoD or Service contractor organizations in these specific areas could also benefit, by knowing what is expected of their personnel and by being able to better select applicants. Another possible transfer involves the use of jet engine mechanic measures by such aircraft manufacturers as Boeing and McDonnell Douglas, as well as the airlines. The list of possible companies that could benefit from job performance measurement technology transfer may be endless.

Without spending time discussing the implications of the Uniform Guidelines on Employee Selection Procedures (1978), suffice it to say that these Guidelines require employers to show a direct link between selection instruments and actual job performance. Interestingly, this is the primary focus of the Joint-Service JPM project. Private sector use of Service-developed instruments and technology will have an immeasurable positive impact on industry's selection, classification, and training decisions.

CONCLUSION

Inter-Service technology transfer has made many promises and provided many payoffs in a short period of time. These payoffs, however, are only a beginning. The transfer of technology, both procedural transfer and transfer-in-kind, has great potential among the Services as well as between the DoD and private industry. Past and current transfer efforts have and will continue to demonstrate cost savings/cost avoidance. Technology transfer is a reality, but to fully realize the multitude of possible payoffs, we must all march to the beat of the same drummer.

REFERENCES

AFM 50-2. (1986, July). Instructional systems development. Washington, DC: Department of the Air Force.

AFM 50-5. (1986, December). USAF formal schools catalog and specialty training standards. Washington, DC: Department of the Air Force.

AFM 66-1. (1983, April). Maintenance management policy. Washington, DC: Department of the Air Force.

AFP 50-58. (1978, July). Handbook for designers of instructional systems. Washington, DC: Department of the Air Force.

AFR 35-2. (1982, July). Occupational analysis program. Washington, DC: Department of the Air Force.

AFR 39-1. (1983, January). Airman classification regulation. Washington, DC: Department of the Air Force.

AFR 50-23. (1982, September). On-the-job training. Washington, DC: Department of the Air Force.

AFR 50-38. (1986, March). Field evaluation of education and training programs. Washington, DC: Department of the Air Force.

ATCR 52-3. (1984, January). Student measurement. Washington, DC: Department of the Air Force.

ATCR 52-12. (1986, September). Training evaluation. Washington, DC: Department of the Air Force.

Alba, P.A. (1986). Transfer of Air Force performance measurement technology to the Navy: Final report. Unpublished manuscript.

Alba, P.A., Dickinson, T.L., & Lipscomb, M.S. (1985). Walk-Through Performance Testing documentation for jet engine mechanics (AFS 426X2). Final report prepared for the Air Force Human Resources Laboratory, Brooks AFB, TX.

Allen, M.J., & Yen, W.M. (1979). Introduction to measurement theory. Monterrey: Brooks/Cole.

Anastasi, A. (1982). Psychological testing (4th ed.). New York: Macmillan.

Baker, H.G., & Blackhurst, J.L. (1986). Inter-Service technology transfer: Performance testing of jet engine mechanics. Proceedings of the 28th Annual Conference of the Military Testing Association. New London, CN: U.S. Coast Guard Academy.

Banks, C.G. (1987, February). Job performance measurement (JPM) technology application to training evaluation (Technical Paper under review). Brooks AFB, TX: Air Force Human Resources Laboratory

Blackhurst, J.L., & Baker, H.G. (1985). Inter-Service transfer of job performance measurement technology. Proceedings of the 27th Annual Conference of the Military Testing Association. San Diego, CA: Navy Personnel Research and Development Center.

Bock, R.D. (1975). Multivariate statistical methods in behavioral research. New York: McGraw-Hill.

Borman, W.C. (1979). Format and training effects on rating accuracy and rating errors. Journal of Applied Psychology, 64, 410-421.

Carson, S.B., Chambers, L.D., & Gosc, R.L. (1984, March). Integrated training system for Air Force on-the-job training: Specification development (AFHRL-TP-83-54, AD-A139 804). Lowry AFB, CO: Training Systems Division, Air Force Human Resources Laboratory.

Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (Ed.), Educational measurement (2nd ed.) (pp. 443-507). Washington, DC: American Council on Education.

Department of Defense (1982). Profile of American youth: 1980 nationwide administration of the Armed Services Vocational Aptitude Battery. Washington, DC: Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs and Logistics).

DeVries, P.B., Jr., Eschenbrenner, A.J., Jr., & Ruck, H.W. (1980, July). Task analysis handbook (AFHRL-TR-79-45(II), AD-A087 711). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Dickinson, T.L., Hedge, J.W., & Ballentine, R.D. (1987, March). Predictability of the Armed Services Vocational Aptitude Battery for the Air Force's Job Performance Measurement System. Paper presented at the DOD/ETS Conference on Job Performance Measurement Technologies, San Diego.

Driskill, W., Mitchell, J., & Ballentine, R.D. (1985, November). Using job performance measures as criterion for evaluating training effectiveness. Unpublished manuscript.

Dunnette, M.D. (1963). A note on the criterion. Journal of Applied Psychology, 47, 317-323.

Ford, J.K., & Wroten, S.P. (1984). Introducing new methods for conducting training evaluation and for linking training evaluation to program redesign. Personnel Psychology, 37, 651-665.

Gibson, R.S., & Orlansky, J. (1986, September). Performance measures for evaluating the effectiveness of maintenance training (IDA Paper P-1922). Washington, DC: Institute for Defense Analysis.

Guion, R.M. (1976). Recruiting, selection and job placement. In M.D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally.

Guion, R.M. (1979). Principles of work sample testing: II. Evaluation of personnel testing programs (ARI TR-79-A9). Alexandria, VA: U.S. Army Research Institute.

Hedge, J.W. (1984, August). The methodology of Walk-Through Performance Testing. Paper presented at the annual meeting of the American Psychological Association, Toronto.

Hedge, J.W., Ballentine, R.D., & Gould, R.B. (1985, October). Examining the link between training evaluation and job performance criterion development. In J.D. Hagman (Ed.), Symposium on the transfer of training to military operational systems. (pp. 308-322). Brussels, Belgium: HQ NATO.

Hedge, J.W., & Teachout, M.S. (1986, November). Job performance measurement: A systematic program of research and development (AFHRL-TP-86-37, AD-A174 175). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Joreskog, K.G. (1969). A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 34, 183-202.

Joreskog, K.G. (1971). Statistical analysis of sets of congeneric tests. Psychometrika, 36, 109-133.

Kavanagh, M.J., Borman, W.C., Hedge, J.W., & Gould, R.B. (1986, February). Job performance measurement classification scheme for validation research in the military (AFHRL-TP-85-51, AD-A164 837). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Klimoski, R.J., & London, M. (1974). Role of the rater in performance appraisal. Journal of Applied Psychology, 59, 445-451.

Landy, F.J. (1986). Stamp collecting versus science. American Psychologist, 41, 1183-1192.

Lawler, E.E., III. (1967). The multitrait-multirater approach to measuring managerial job performance. Journal of Applied Psychology, 51, 369-381.

Lennon, R.T. (1956). Assumptions underlying the use of content validity. Education Measurement, 16, 294-304.

Lipscomb, M.S. (1984, August). A task-level domain sampling strategy: A content-valid approach. Paper presented at the annual meeting of the American Psychological Association, Toronto.

Mifflin, T.L., & Verna, S.M. (1977). A method to correct correlation coefficients for the effects of multiple curtailment (CRC-336). Arlington, VA: Marine Corps Operations Analysis Group, Center for Naval Analyses.

Mitchell, J.L., & Driskill, W.E. (1979, October). Variance within occupational fields: Job analysis versus occupational analysis. Proceedings of the 21st Annual Conference of the Military Testing Association, (pp. 259-268). San Diego, CA: Navy Personnel Research and Development Center.

Pennell, R., Harris, D., & Schwille, J. (1976, October). Appraisal of Air Force training course field evaluation system (AFHRL-TR-76-63, AD-A035 641). Lowry AFB, CO: Technical Training Division, Air Force Human Resources Laboratory.

Solomon, H. (1986, March). Economic issues in cost-effectiveness analysis of military skill training (IDA Paper P-1897). Washington, DC: Institute for Defense Analysis.

Uniform guidelines on employee selection procedures. (1978). Federal Register, 43, 38290-38315.

Wilson, C.L. (1962). On-the-job and operational criteria. In R. Glaser (Ed.), Training research and education. Pittsburgh: University of Pittsburgh Press.

END

DATE

FILMED

9-88

DTIC